Detecting Atypical Examples of Known Domain Types by Sequence Similarity Searching: The SBASE Domain Library Approach

Somdutta Dhir[1], Mircea Pacurar[1], Dino Franklin[2], Zoltán Gáspári[3], Attila Kertész-Farkas[1], András Kocsor[4], Frank Eisenhaber[5,6,7] and Sándor Pongor[1]*

[1]*Protein Structure and Bioinformatics, ICGEB, Trieste, Italy,* [2]*Department of Computer Science, Federal University of Uberlândia (UFU), Brazil,* [3]*Laboratory of Structural Chemistry and Biology, Institute of Chemistry, Eötvös Loránd University, Budapest, Hungary,* [4]*Research Group on Applied Intelligence NPC, Szeged, Hungary,* [5]*Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), 30 Biopolis Street, Singapore,* [6]*Department of Biological Sciences (DBS), National University of Singapore (NUS), 8 Medical Drive, Singapore,* [7]*School of Computer Engineering (SCE), Nanyang Technological University (NTU), 50 Nanyang Drive, Singapore*

Abstract: SBASE is a project initiated to detect known domain types and predicting domain architectures using sequence similarity searching (Simon *et al.*, *Protein Seq Data Anal*, 5: 39-42, 1992, Pongor *et al.*, Nucl. Acids. Res. 21:3111-3115, 1992). The current approach uses a curated collection of domain sequences – the SBASE domain library – and standard similarity search algorithms, followed by postprocessing which is based on a simple statistics of the domain similarity network (http://hydra.icgeb.trieste.it/sbase/). It is especially useful in detecting rare, atypical examples of known domain types which are sometimes missed even by more sophisticated methodologies. This approach does not require multiple alignment or machine learning techniques, and can be a useful complement to other domain detection methodologies. This article gives an overview of the project history as well as of the concepts and principles developed within this the project.

Keywords: SBASE domain library, sequence similarity searching, protein domain prediction, atypical domain detection.

INTRODUCTION

Classification of proteins is a fundamental technique in computational genomics. Prediction of known domain types within a protein sequence can be regarded as a special subcase of protein classification. Namely, while general classification methods assign a general, global descriptor (annotation) to an entire protein, domain prediction methods use similar computational techniques to assign local descriptors to a specific segment of a protein sequence. Generally speaking, domains are the structural and functional building blocks of proteins that can exist, evolve and function independently of the rest of the protein chain. While globular domains have a defined 3D structure and their sequences can be readily identified with bioinformatics methods, non-globular segment have biased composition and pose problems in similarity searching. We start our overview by placing protein domain prediction into the framework of protein classification.

General methods of protein classification fall into four broad categories. i) Pair-wise comparison methods work by comparing an unknown object (protein sequence or structure) with members of an *a priori* classified database of protein objects. The results are ranked according to the similarities and the strongest similarities are evaluated in terms of biological or statistical significance, after which a query is assigned to the class of the most similar object. ii) Generative models are based on consensus (or aggregate) descriptions of protein groups. Methods for preparing consensus descriptions include regular expressions, frequency matrices, profiles and Hidden Markov Models. The unknown query is then compared to a collection of generative models and the strongest similarities are evaluated and used to assign the protein to the given class. iii) Discriminative models seek to determine a boundary between a class (positive group) and its immediate similarity neighborhood (negative classes). Such boundaries are established with learning algorithms, among which kernel methods, in particular support vector machines (SVMs) are extremely popular. iv) Network models use a graph-like representation in which proteins are the nodes and similarities are the (weighted) edges. Such a network can be evaluated by simple local statistics or by propagation algorithms such as the PageRank algorithm [1] used in the Google search engine that was successfully applied later to protein similarity searching [2].

The overall development trend can be best traced by how the background similarities are incorporated into the classification scheme. On the one extreme, supervised classification schemes assume *a priori* knowledge on all protein classes i.e. a full knowledge of the background similarities, while on the other extreme, unsupervised classification assumes none. Protein domain prediction methods we are concerned with, fall into the broad mathematical category of supervised classification, but they differ widely in the form and amount of background knowledge used in the analysis. For example, when we accept pairwise similarities only above a given score value, we concentrate our knowledge of the background into a single number, a database-wide threshold value. Or, when we use generative models, such as HMMs, we assume knowledge on all protein classes. If we train

*Address correspondence to this author at the ICGEB, Padriciano 99, 34149 Trieste, Italy; Tel: +39-040-3757300; Fax: +39-040-226555;
E-mail: pongor@icgeb.org

1389-2037/10 $55.00+.00

SVMs for the protein groups, we assume knowledge on all classes as well as their neighborhoods. In the case of network models, the knowledge is implicit to the similarity network. Finally, protein classification methods differ in their needs for human intervention. Generative models based on multiple alignment (MA) need human experts to look at the alignments, which add to the overheads of database maintenance.

We can now place the SBASE approach in the hierarchy of protein classification methods. SBASE is a project established for detecting protein domain types based on the sequence. It uses a database of domain sequences collected from curated primary sequence databases – the SBASE domain library – as well as a set of similarity search algorithms that operate either on this collection or on the primary sequence databases themselves. SBASE search algorithms use pairwise similarities for domain detection, without MAs. Instead of a single, database-wide threshold, SBASE uses individual threshold values for the various protein classes. In addition, the threshold values are applied to network parameters such as the number of links between proteins, so the properties of the similarity network are also taken into account. In summary, SBASE tries to build on various parameters of the background similarities that can be simply extracted from an all vs. all database comparison. This makes it possible to have a system that can be maintained with minimal human intervention and minimal computer resources. As opposed to generative models used by such comprehensive collections as the Interpro project [3], SBASE is a database-driven approach that facilitates detection of atypical examples Fig. (**1**). From among the BLAST-based, database-driven systems that are more related to SBASE, PRODOM uses automatically generated domain families [4], as opposed to the curation-based classification scheme adopted in SBASE. Finally, the SYSTERS database developed by the Vingron group is based on an iterated researching algorithm on entire protein sequences that makes it possible to assign a sequence to a sequence cluster [5].

The rest of this paper is structured as follows. We first review the project history and the development of the main ideas. Next, we describe the SBASE protein domain sequence library and its use in domain architecture prediction. Subsequently we describe the issues of performance checking and database maintenance. The review finishes with a summary of conclusions and future trends in which we place the approach within the broader context of segment-based annotation techniques. The review is accompanied by a Glossary of terms and concepts used, and partly developed within the SBASE project (Appendix).

MOTIVATION, PROJECT HISTORY

The generative models used for protein identification in the early 90's were based on MAs. The motivation behind the SBASE project was to use pairwise alignments directly for finding known domains, without the necessity to construct and curate MAs and /or generative models. The reason was three-fold: i) Generative models (such as consensus sequences, regular expressions, sequence profiles, hidden Markov models etc) and MAs are difficult to maintain in an automated way. The process requires human expertise and careful judgment hence it can hardly keep pace with the flow of new genome data. ii) Generative models unify the description of individual sequences included so they are necessarily biased towards an average description. This makes the discovery of atypical proteins difficult, which is a major problem since new genomes inevitably contain novel variants of the known proteins. iii) There are domain types for which it is not easy to develop consensus representations because of weak similarities. It was speculated that an annotated protein database record contains the domain architecture and the sequence of a database entry, so an alignment against such a record should give direct indications on domain homologies. The knowledge base of such a system is the annotated sequence database itself, the task was to design tools that can combine sequence alignment scores with the annotation information (domain architecture, "feature table"). From this point of view, domain similarity search is a two-fold sorting exercise: the highest scoring similarity regions (such as BLAST HSPs (High-scoring Segment Pairs)) have to be sorted by score as well mapped to the query sequence. Two principles were considered Fig. (**2**):

1) The domain library approach consists in building a domain sequence collection using annotated sequence segments collected from annotated sequence databases. Search against this database would directly give domain similarities which can be easily turned into domain architecture cartoons. The advantage of this approach is that one can use any similarity search program [6, 7].

2) The FTHOM approach was named after the post-processor program built for the purpose. This approach consists of searching a database of annotated proteins (such as Uniprot) and evaluating the domain similarities by comparing the similarity regions of the output with the feature table (domain architecture) given in the annotation part of the database record. The advantage of this approach is that one does not need a separate domain collection, however it does not provide boundaries between domains of the same type [8, 9].

Both approaches depend on two factors: i) The quality of domain annotations in primary databases. In the early databases, annotations were given in highly overlapping fashion, so regions, repeats, domains, biased composition regions were all part of the feature table. One had to select which of these should be considered in the domain prediction process. ii) The approach used by the search programs. Each search program defines similarity regions in a slightly different manner, some only give the highest scoring regions, others, like BLAST [10] give several local alignments. Naturally, the speed of the search is a major consideration, and since 1998 BLAST has become the main search tool.

The first published result of the project was a search against a sequence collection of known domains using the FastDB program [6, 7]. As far as we are aware this was the only publicly available domain sequence collection at the time (Uniprot and annual releases were published in the subsequent years) [11-21].

An important step of the development was to include the entire network of sequence similarities into the evaluation process [22, 23]. So query sequences were evaluated not

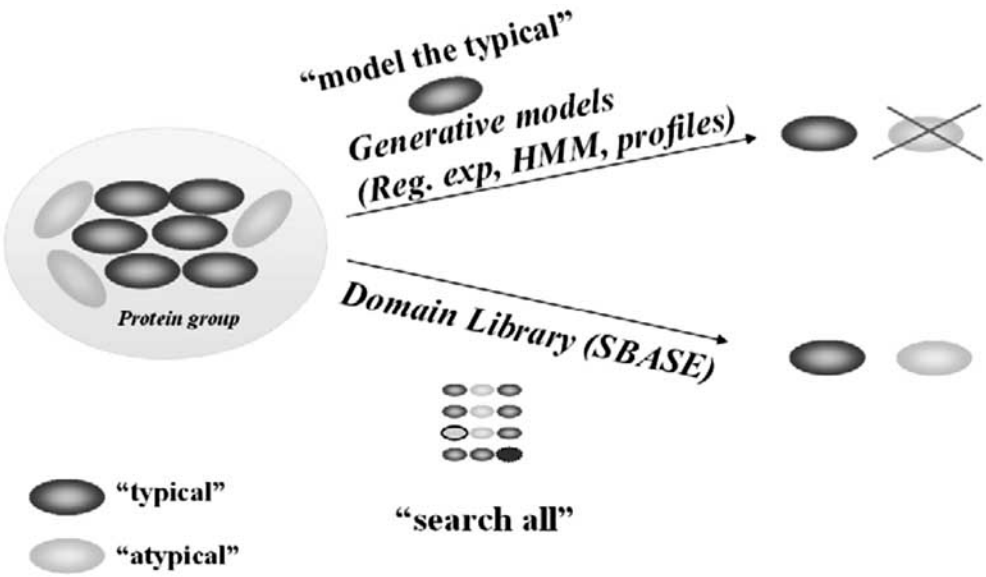

Fig. (1). Database-driven approach. Database driven approaches give an equal weight to atypical and typical domain sequences which are sometimes overlooked as "noise" when the generative models are built on an ensemble of exemplars.

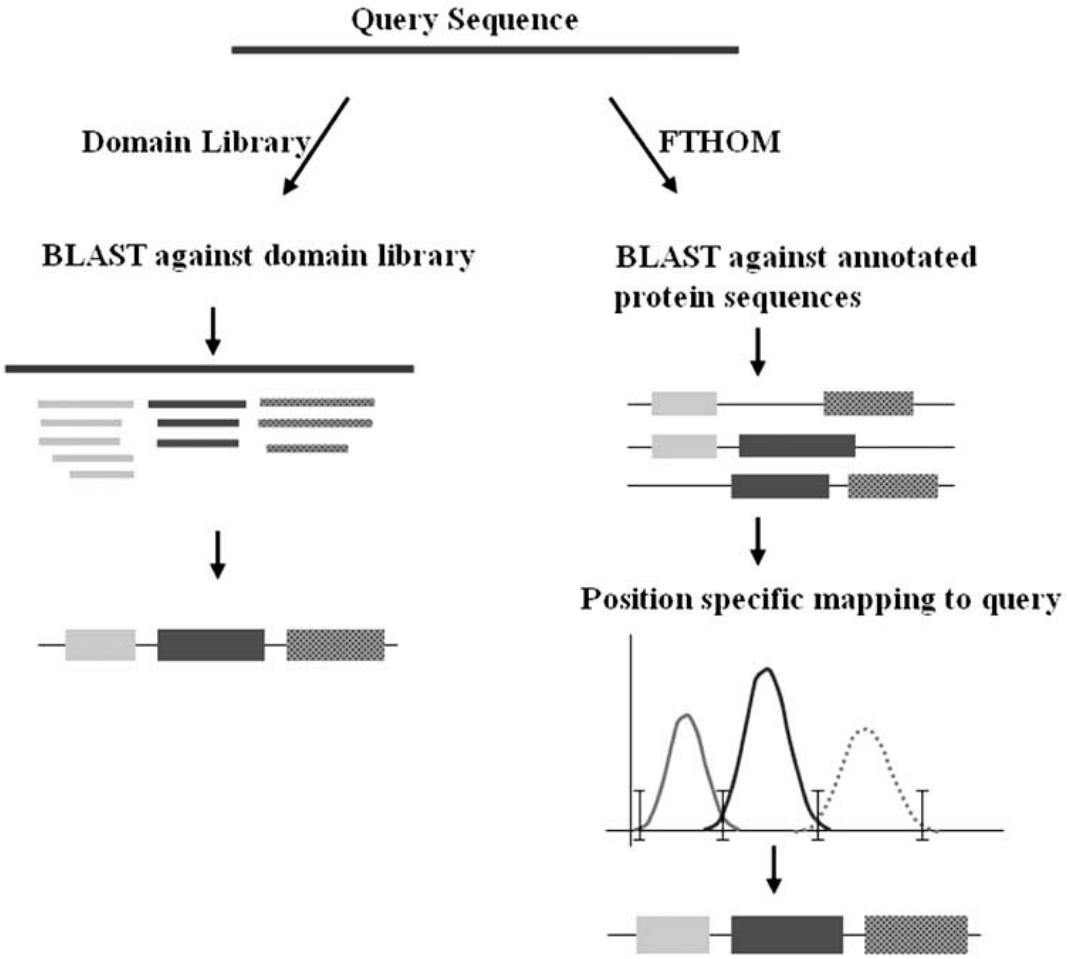

Fig. (2). Database-driven segment annotation. Two principles of database-driven segment annotation. **(a)** In the domain library approach, the query is compared with a database of annotated protein segments and the regions of similarities are mapped to the query sequence to give a final domain assignment. **(b)** In the FTHOM (feature homology search) approach, the query is compared with a database of fully annotated proteins, and the alignments are mapped in a position-specific manner to the query sequence. This gives rise to a series of sequence plots that are then computationally processed to give a final domain assignment.

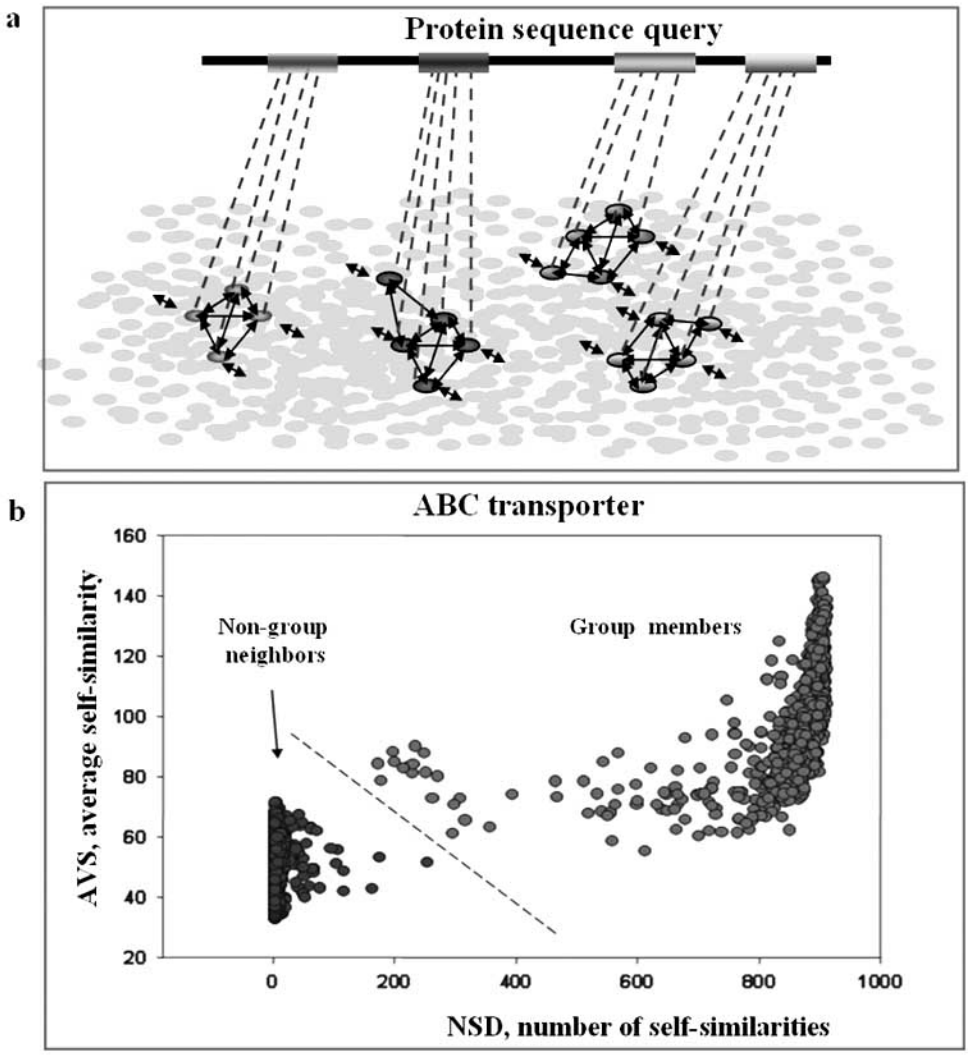

Fig. (3). The similarity network approach. **(a)** The library of known domain sequences is compared to itself with a program so as to give a network of similarities. Based on the known domain classes, within group and between group similarities ("self" and "non-self" respectively) are identified. When a protein query is compared with the domain library, assignment is based comparing the query vs. group similarities with the self and non-self similarities. **(b)** The neighborhood of each domain group is described as a vector of local network parameters [22]. A 2D plot shows the separation of self and non-self group members around the ABC transporter group. Such low-dimensional descriptions are easily amenable to machine learning applications [9, 13].

only with a simple database search, but an all vs. all comparison of the database against itself was also used as background knowledge. In this perspective, a similarity search provides links between a query and its immediate neighborhood. The similarity network (i.e. the network of all known domains) contains all neighborhoods that appear as more or less densely connected regions within the network, where the nodes are domains and the edges are significant similarities weighted by a parameter such as the BLAST score. When a new sequence is compared with a database represented as a similarity network, a similarity region (e.g. BLAST HSP) will be classified to the closest neighborhood, provided the similarities exceed certain thresholds Fig. (**3**). This approach is a discriminative model where the important parameters (e.g. thresholds of within-group and between-group similarities define the boundaries of a neighborhood) are simply extracted from the all *vs*. all comparisons. The similarity network can thus be regarded as the memory of the system that is built continuously as new domains become known.

The evaluation of the results was based on sequence similarity, even though for "difficult domain types" that typically constitute 5-10 percent of the known domain types, machine learning algorithms were also used such as neural networks [9] and SVMs [13]. Machine learning applications for protein classification usually rely on a description of the protein structure or sequence that can be turned into a vector, suitable for training. The distinctive feature of the neural network and SVM applications used in the SBASE project was that they used vectorial descriptions of the sequences that were derived from a similarity with respect to protein groups. This is a low dimensional, aggregate description that builds on the salient features of the similarity space.

From the beginning, SBASE was conceived as an academic pilot project carried out by a small research group. It became clear however that the collection could not keep pace with the amount of data, so from 2006 we replaced the original domain collection by a curated subset of INTERPRO collection [3] that we complement with established domain types from other sources (PFAM [24], SMART [25], SwissProt annotations [26] etc.). At present the SBASE project concentrates on testing various experimental sequence similarity search principles that are incorporated into the web servers maintained at ICGEB Trieste.

THE SBASE DOMAIN LIBRARY

SBASE was conceived as a database of annotated protein sequence segments. In the early years, the sequence segment definitions were taken from databases such as SwissProt, PIR, and additional entries were collected from translations of such as EMBL [27] and GenBank [28] as well as from research articles recommended by colleagues, on an ad hoc basis. As the database annotations contained not only protein domains but also biased composition regions, repeats, signal peptides, etc. – including regions that were identified by computation –, some of these were also included into the collection. The early releases (1.0 to 9.0) are available at ftp://ftp.icgeb.trieste.it/pub/SBASE/. The early releases were comprehensive, in the sense that we tried to include, whenever possible, all known instances for all domain types.

From 2003 the releases were accessible only via the web interface and from 2006 we only included (mostly globular) protein domain types annotated within the INTERPRO collection, so we did not include signal peptides, transmembrane regions, biased composition regions, etc. This allowed us to bring down the collection size to around half a million. From about 2005 even this reduced set of domain types contained more than a million entries so we started to filter the database according to 90 percent identity. This allowed us to bring down the database size again to a manageable size while keeping representatives of the atypical domains. The current size of the SBASE collection is ~736 thousand domain sequences. The statistics of the protein groups includes phylogenetic coverage, references to primary sources as well as a list of calculated physicochemical and other measures Fig. (**4**).

Domain Prediction Using the SBASE Library

The current search engine behind SBASE is BLAST. Given a protein sequence query, the system will first identify signal peptide and transmembrane regions, then submit the corresponding sequence to BLAST that will compare it with the SBASE collection. The domain similarities exceeding group-wise thresholds are mapped on the query sequence according to a greedy algorithm that proceeds in descending order of the similarity scores. An output example is shown in Fig. (**5**). The strength of the method is its sensitivity towards the known atypical examples of protein domains that are often missed even by more sophisticated methods such as generative models see Fig. (**1**). An example of this are the highly variable HEAT repeats whose instances easily detected by human annotators but are often missed by generative models. At the same time, SBASE based either on BLAST, or by Smith-Waterman, detects most of these repeats Fig. (**6**). With this we do not claim that similarity based segment annotation is more sensitive than other techniques, just that it can be a useful complement to other, more sophisticated methods.

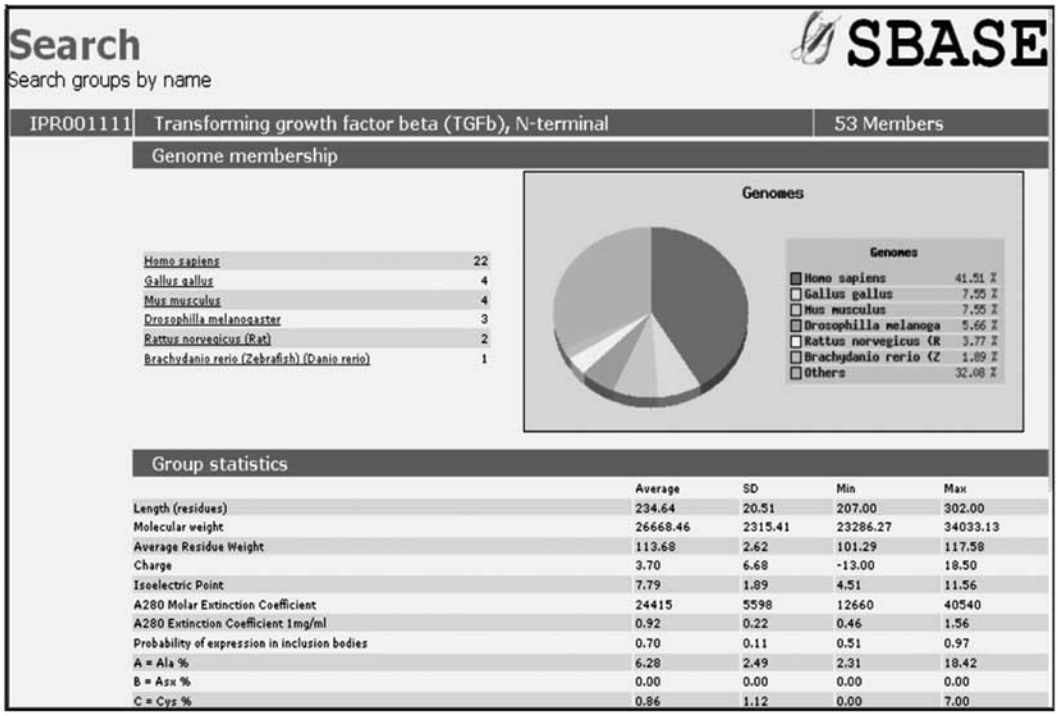

Fig. (4). Statistical summary of a protein group of SBASE. The summary includes the phylogenetic distribution of the protein group (domain type), the reference to the sequence sources as well as a calculated statistics of sequence composition features.

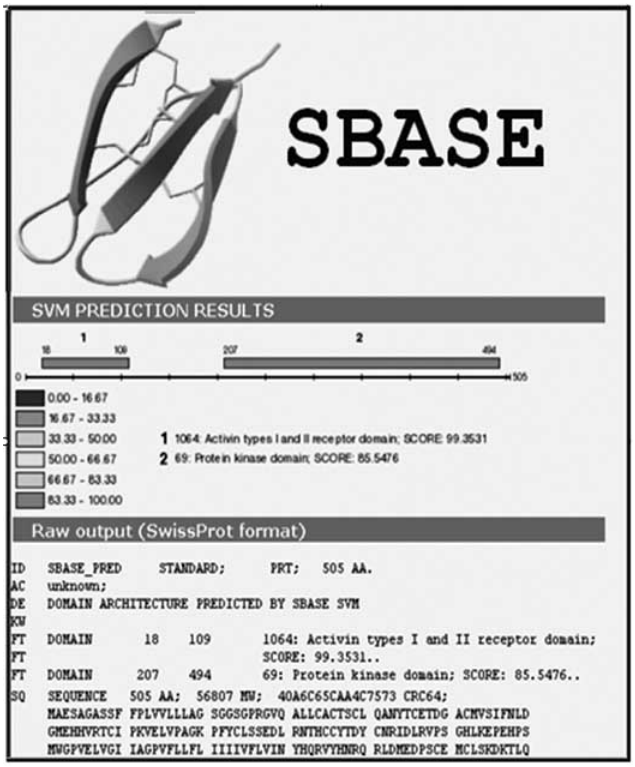

Fig. (5). Search output of the SBASE domain predictor.

Performance Checking

The performance of domain predictors can be evaluated according to the known principles of classification, i.e. using the well known scheme of true positives (tp), false positives (fp), true negatives (tn), false negatives (fn). Characterizing the performance of domain prediction according to this scheme is only seemingly easy, because of a few theoretical and statistical issues.

The discrepancies between predicted and known domain architectures can fall into different categories that can not be easily summarized by a single performance measure. The strategy used for developing SBASE was based on following a hierarchy of statements as follows:

a) Presence-absence level: a protein that has a certain domain type and is predicted to have at least one is a tp. So if we predict one single Ig-like domain in Titin (152 Ig-like domains) it counts already as a positive hit.

b) Composition level: A protein that has *n* number of copies of a certain domain type should be predicted as having at least the same number of domains in order to qualify as tp. If less domain copies are predicted, the prediction is incorrect.

c) Architecture level: A domain should be predicted with the correct boundaries, say within 5 amino acids from those known. In this case, if any of the domains of a given type have boundaries outside the given range, the prediction is incorrect.

Of these, level a) is very permissive, all methods will perform equally well at this level. On the other hand, it is not easy to find a gold standard for levels b) and c), because the known collections differ in many details, the number of domains often differs, atypical domains are frequently missing, etc.

The statistical problems are related to the structure of the protein similarity space which is shaped by domain evolu-

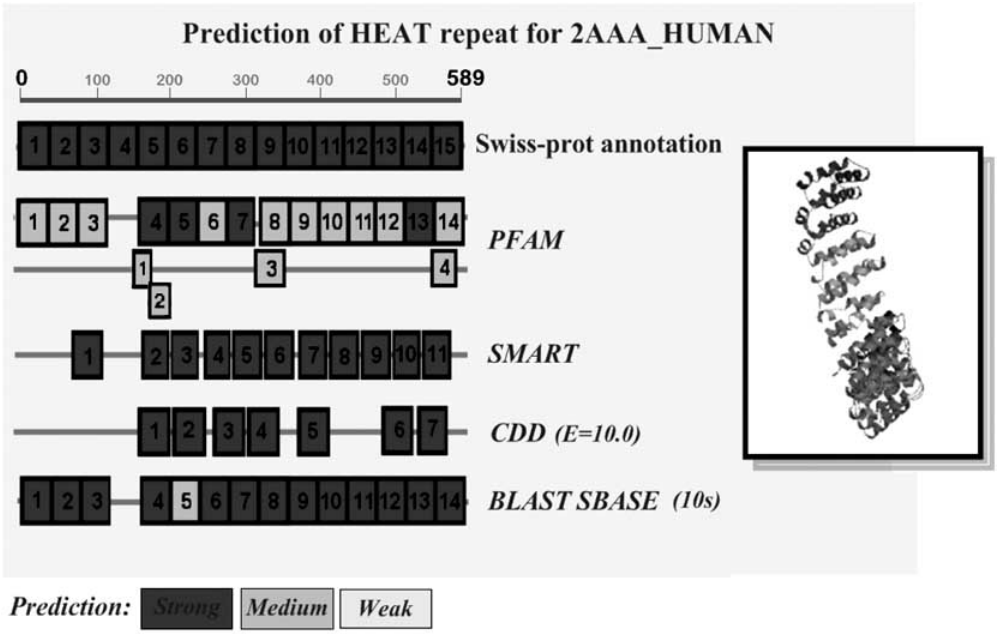

Fig. (6). Prediction of HEAT repeats by various methods. HEAT repeats are ~40 residue long helical hairpin structures of highly variable sequence that assemble to larger structures. The regulatory subunit of human serine phosphatase 2A has 15 such repeats annotated in Swiss-Prot (2AAA_HUMAN). This type of repeat is difficult to detect by domain prediction servers using the default settings (PFAM, *lower line*). Some servers, such as CDD detect a few repeats at higher sensitivity parameter settings, but none at lower sensitivity settings. BLAST SBASE detects all repeats at standard settings, except the 4th repeat which is not detected by any of the servers.

tion. Namely, the real test of a domain predictor is to recognize novel variants of known domain type. So if we want to check how well a certain domain in a given parent protein is predicted, we have to make sure that either a) the parent protein or b) the whole proteome of the parent protein is excluded from the construction of the predictor. A similar principle was applied by the groups of D. Haussler [29] and W. Noble [30] who trained predictors on protein families of a given superfamily (training set) and tested it on members of a protein family not included in the training. Applying this principle to machine learning algorithms is laborious, since in order to systematically exclude the homologues of test queries from the training set, one has to build a new classifier (new generative or discriminative model) for each test case. Methods using MAs pose extra difficulties since, at least in principle the MAs should be rebuilt (and manually checked) for each test case. For instance, it is not sufficient to delete some of the proteins from an existing MA because the gap structure of the MA will continue to contain information on the deleted proteins (i.e. on the test group). These statistical difficulties can be relatively easily taken care of with systems based on pairwise similarities such as SBASE since the various threshold values that depend on the training set can be simply and automatically recalculated for each test case and no time-consuming learning phase is involved. So one can easily ask such questions as, e.g. "How would the system perform if we were to discover this organism right now?", or "How could we predict the first eukaryotic genome based on prokaryotic genomes only"?

Testing and comparing domain predictors is difficult since it requires standardization of i) datasets and classification tasks; ii) standardized sequence/structure comparison methods; iii) classification algorithms; and iv) a validation protocol. On the other hand, the published results are often based on different and sometimes obsolete datasets and program versions and without sufficient detail necessary to reproduce the calculation procedure. In order to alleviate these problems we developed a benchmark collection of sequences [31, 32] In this system we use ROC (receiver operating characteristics) analysis for testing predictor performance [33].

A separate problem is related to the heterogeneity of the protein groups that range in size from 2-3 members to several thousand members. One problem is the overwhelmingly large size of the negative dataset, which is often treated by truncating the top lists [34]. This approach can help one to compare the performance of predictors on the same group but does not allow one to compare groups of different sizes using the same predictor. As the latter task is crucial for pinpointing difficult groups in a domain database, we developed the method of balanced ROC analysis, where the size of the negative group is standardized based on the size of the positive group [35].

Maintainability Issues

Maintenance of a domain sequence database is more than simply collecting sequences. Today it also includes maintenance of group descriptions that include information of functional roles, phylogenetic coverage, 3D structure etc which is a task that only large groups of professional annotators can cope with. The organization and visualization of the data is also becoming a separate field and some sites offer highly rich and complex views on domain architectures. The times where databases were striving for complete coverage of sequence data have apparently passed, with the vast number of ongoing sequencing projects not even the largest collections can realistically hope for comprehensive coverage of known sequences.

The strategy chosen within the SBASE project is to reach good *predictive coverage*, i.e. we try to include a) newly published domain sequence groups, as well as b) new members of established groups provided they are more than 10% different from the known members of the group. All this keeps the maintenance overheads quite low while allows for a predictive coverage that compares well with other, sophisticated search programs.

CONCLUSIONS AND FUTURE TRENDS

Against a fast evolving background of sequence collections and prediction methods, it is difficult to predict the future of a small academic project like SBASE. On the other hand SBASE can easily cope with atypical domains so it is a useful tool to complement other methods of domain prediction.

From the methodological point of view, the approach of SBASE is part of a larger group of methods termed segment-based sequence annotation [36]. Namely, SBASE assigns discrete domain annotations to protein segments based on a particular method, sequence similarity searching. There are many other computational tools that can assign a variety of annotations based on different principles that are either discreet or parametric (quantitative) in character Fig. (**7**). Discrete methods assign a feature on a yes/no basis, while a parametric method provides a quantitative measure that can be plotted along the sequence. In such a plot the putative segments appear as peaks. Summary feature descriptions Fig. (**7**, bottom) are reminiscent of the familiar domain architecture cartoons, the important difference being that computationally generated segments can be overlapping.

A genome annotation system may use a battery of these methods to assign segments. For example ANNOTATOR is an annotational pipeline [37] that uses over 20 of the most useful annotation algorithms (Table **1**) covering the first two steps of segment- based sequence analysis [38] that have proven to be indispensable in daily sequence analytic work for functional discovery [39-41]. Of particular value is the inclusion of predictors for a number of post-translational modifications [42-47] as well as targeting signals [48, 49]. ANNOTATOR is an environment designed to aid function prediction for uncharacterized protein targets that are difficult because of the absence of closely related, well characterized homologues. Therefore, it needs to include as many as possible prediction tools that recognize various sequence pattern-function relationships. We think that including a protocol of domain detection based on SBASE would further enhance the predictive power of the ANNOTATOR environment.

In addition, there are a number of technical improvements that can increase the competitiveness of database-driven methods. The GPU-based version of the Smith-Waterman algorithm produces run times approaching those

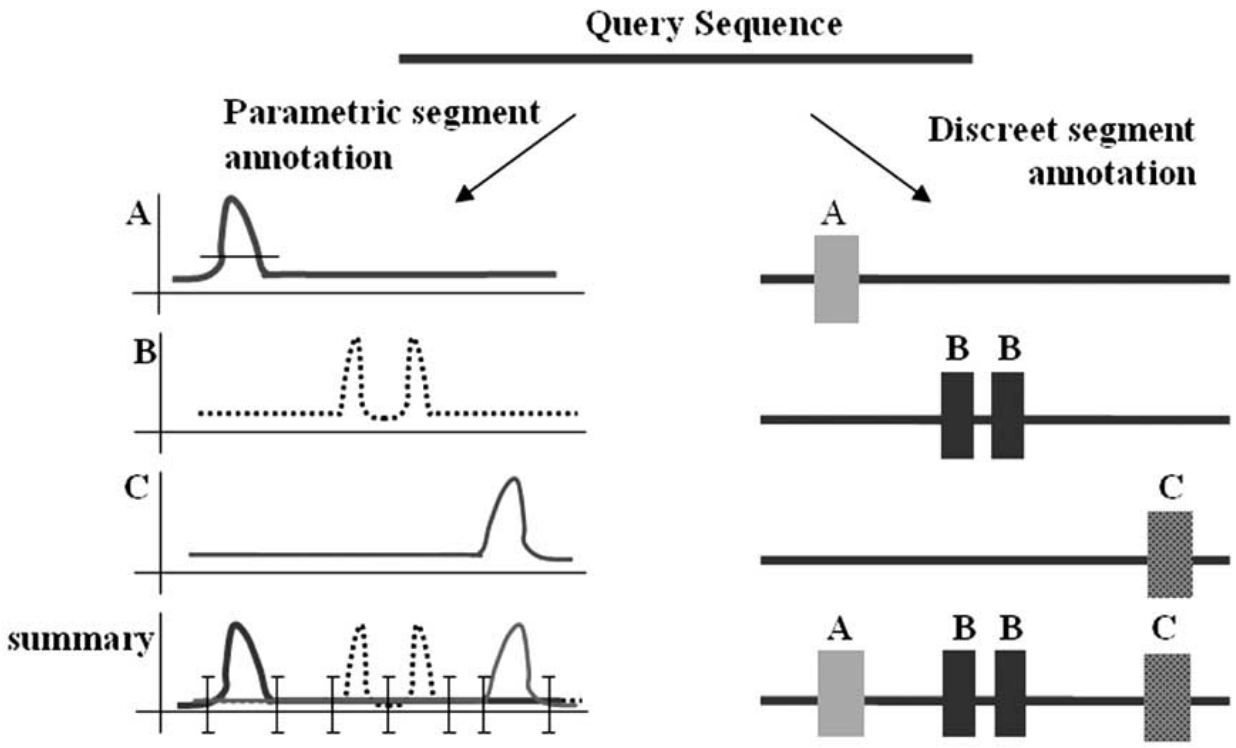

Fig. (7). Segment-based annotation uses either a quantitative parameter (left) or a discrete yes-no decision (right) to assign annotations to sequences. In the parametric case, we have as many sequence plots as there are segment annotation types. A sequence plot can be transformed to segments, for instance by applying a threshold value (as shown under A, left). Summary descriptions are reminescent to the domain architecture cartoons. The domain similarity plot of FTHOM (Fig. (**2**)) is in fact a parametric domain similarity summary.

Table 1. Algorithms Used in the Automated Annotation Pipeline ANNOTATOR [37]

Algorithm	Description	Type*	Reference
CAST	Algorithm for low-complexity region (LCR) detection and selective masking	1	[50]
IUPred	Prediction method for recognizing ordered and intrinsically unstructured/disordered regions in proteins	1,2	[51]
SAPS	Statistical analysis of protein sequences with respect to amino acid composition and simple sequence motifs	1,2	[52]
SEG	Prediction of low complexity regions	1	[53]
Big-∏	Prediction of protein GPI lipid anchor cleavage sites	4	[42-44]
NMT	Prediction of N-terminal N-myristoylation of proteins	4	[45, 46]
PrePS – FT	Farnesylation prediction	4	[47]
PrePS – GGT1	Geranylgeranylation prediction	4	[47]
PrePS – GGT2	Rab geranylgeranylation Prediction	4	[47]
PeroPS/PTS1	Prediction of peroxisomal targeting signal 1	4	[48, 49]
DAS-TMfilter	Prediction of transmembrane regions	1,5	[54]
HMMTOP	Transmembrane topology prediction using Hidden Markov models	1	[55]
PHOBIUS	Combined transmembrane topology and signal peptide predictor	4	[56]
TMHMM	Transmembrane helix predictor	1,2	[57]
IMP-COIL	Prediction of coiled-coil regions, modified implementation of the algorithm Lupas et al. by F. Eisenhaber	1,2	[58]

(Table 1) contd….

Algorithm	Description	Type*	Reference
PROSITE	Pattern search in the PROSITE database	3	[59]
PROSITE-Profile	Profile search in the PROSITE database	3	[59]
HMMER	Profile Hidden Markov Models	3	[60]
IMPALA	Tool to compare a query sequence against a library of position-specific scoring matrices	3	[61]
RPS-BLAST against CDD	Reverse-position-specific BLAST against the Conserved Domain Database (CDD)	3	[62]

*Types: 1=calculaton on the fly, 2 = based on statistically derived tabulated values, 3 = generative model, 4 = discriminative model, 5= calculation on a database.

of the BLAST algorithm on CPUs. The GPU-based alignment programs of Manavski and Valle [63, 64] has been adapted to the SBASE system and seems to improve the prediction of atypical domains. Prediction quality can further improve by including a jury of predictors that use alignment parameters which can be simply extracted from the search results [65]. Taken together, similarity based detection methods are useful complements to other, more sophisticated domain detection methodologies and their simplicity and versatility makes them suitable to solve annotation tasks, especially in the view of increased computational speed of currently appearing computer technologies.

ACKNOWLEDGEMENTS

The authors are indebted to all previous coworkers who helped the earlier phases of the project, especially to János Murvai for his fundamental role in developing the algorithms and programs, as well as to Hedvig Hegyi (FTHOM) and Kristian Vlahoviček (first WWW server applications). The work at ELTE was supported by grants from ICGEB (CRP/HUN08-03), the Hungarian Scientific Research Fund (OTKA F68079 and K72973), as well as a János Bolyai Research Fellowship and a FEBS Short-term Fellowship to Z.G.

APPENDIX

Glossary of Terms and Concepts

Similarity Groups

The similarity group is a group of molecules that share a well-defined type of similarity [66]. This similarity can be local or global, structural or functional etc. Biologically important similarity groups, such as those of protein domains share global structural similarities, as all group members are characterized by a common sequence-description or a common fold-description. On the other hand, they are not necessarily similar in the functional sense. Sequence similarity groups can be best pictured as a weighted graph where nodes are protein sequences and edges are the similarities between them, weighted by similarity score. Sequence pairs with similarity scores below a certain threshold are not connected.

Within-Group and between-Group Similarities

These terms denote similarities between members of the same or of different protein domain groups, respectively. In the ideal case, members of a given group are connected only with themselves, so "well-separated" groups have no between-group similarities. On the other hand, certain "overlapping" or "connected" groups share a significant number of similarities. Such cases can lead to erroneous annotations if only pairwise similarities are considered, however the network-based scoring outlined in Fig. (**3b**) is often successful.

Descriptors, Local and Global

The annotation of proteins can be pictured as attaching descriptors – pieces of information – to a database entry. The annotation part of sequence records contains descriptors that refer to the entire protein, such as the ID, the accession number, the function, etc. These are global descriptors. Other descriptors, such as those in the feature table (domains, active sites, mutations, etc.) refer to a well defined part of the protein, these are local descriptors.

Database-Wide, Group-Specific and Element-Specific Thresholds

When pairwise sequence comparison methods are used for assigning query proteins to functional classes, it is customary to apply threshold values. For instance, it is customary to disregard BLAST similarities if the similarity score is below an arbitrary value (e.g. 40). This is a database-wide threshold since it is uniformly applied to pairwise comparisons between a query and all entries of the database. If the database is pre-classified and the sequence classes are known, one can separately evaluate the within-group similarities for all protein groups, so one can establish group-specific thresholds for each group within the database. In SBASE, we use a set of such group-specific thresholds Fig**. (3b**).

ROC Analysis, Elementwise and Groupwise, Balanced ROC

ROC (ëreceiver operating characteristicsí) analysis is a visual as well as numerical method used for assessing the performance of classification algorithms [33]. The ROC curve is a graphical plot of the sensitivity vs. (1 − specificity) for a binary classifier system as its discrimination threshold is varied. Any ranked list of similarities between a query and members of a preclassified database is suitable to calculate a ROC curve. The performance measure is the integral of this curve (the so-called AUC value) which is 0.5 for random classifiers and 1.00 for a perfect classifier (no false positives

or false negatives). In group-wise comparison, one prepares a single ranked list, in which the members of the positive and negative test group are ranked according to their similarity to the + train group. In the element-wise scenario, one builds a separate ROC curve for the top lists individually taken for all members of the + test group. In protein classification due to the excessive size of the negative class, truncating the top lists after a certain number of false positives [34] makes it possible to compare classifiers on the same protein group. Comparing protein groups among themselves – a crucial problem in finding problematic groups in databases – is possible by truncating the top list at a size proportional to the positive training group [35].

Structured and Unstructured Representations

All molecular structures, including biological sequences can be pictured as graphs consisting of entities and relationships [67]. We use the term "structured descriptions" for those descriptions that contain both entities and relationships. Protein 3-D structures and sequences are such descriptions even though the relationships are not explicitly included in the actual descriptions found in databases. If a description contains only entities or only relationships, we term it an "unstructured description". Examples include amino acid composition (only entities) and Cα distance-distributions (only relationships).

Supervised Cross-Validation

Cross-validation is a technique wherein a classification performance measure – such as ROC AUC – is determined on several or many test and train groups selected from the same set of data. In traditional cross-validation, the test and train groups are selected in a random way. Such tests may not give reliable estimates on how an algorithm will generalize to novel, distantly related subtypes of the known protein classes. The generalization property of a given classifier algorithm can be better estimated by supervised cross-validation [31], when test and train sets are selected according to the known subtypes within each group of database. Hierarchical classification trees of protein categories (such as used in CATH [68], SCOP [69] or COG [70]) provide a simple and general framework for designing supervised cross-validation strategies and benchmark datasets for protein classification [31, 32].

Detecting atypical examples of known domain types by sequence similarity searching: The SBASE domain library approach.

REFERENCES

[1] Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank citation ranking:bringing order to the web*. Technical report, Stanford Digital Library Technologies Project, **1998**.

[2] Weston, J.; Elisseeff, A.; Zhou, D.; Leslie, C.S.; Noble, W.S. Protein ranking: from local to global structure in the protein similarity network. *Proc. Natl. Acad. Sci. USA*, **2004**, *101*(17), 6559-6563.

[3] Hunter, S.; Apweiler, R.; Attwood, T.K.; Bairoch, A.; Bateman, A.; Binns, D.; Bork, P.; Das, U.; Daugherty, L.; Duquenne, L.; Finn, R.D.; Gough, J.; Haft, D.; Hulo, N.; Kahn, D.; Kelly, E.; Laugraud, A.; Letunic, I.; Lonsdale, D.; Lopez, R.; Madera, M.; Maslen, J.; McAnulla, C.; McDowall, J.; Mistry, J.; Mitchell, A.; Mulder, N.; Natale, D.; Orengo, C.; Quinn, A.F.; Selengut, J.D.; Sigrist, C.J.; Thimma, M.; Thomas, P.D.; Valentin, F.; Wilson, D.; Wu, C.H.; Yeats, C. InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **2009**, *37*, D211-D215.

[4] Bru, C.; Courcelle, E.; Carrere, S.; Beausse, Y.; Dalmar, S.; Kahn, D. The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.*, **2005**, *33*, D212-D215.

[5] Krause, A.; Stoye, J.; Vingron, M. Large scale hierarchical clustering of protein sequences. *BMC Bioinformatics*, **2005**, *6*, 15.

[6] Simon, G.; Paladini, R.; Tisminetzky, S.; Cserzo, M.; Hatsagi, Z.; Tossi, A.; Pongor, S. Improved detection of homology in distantly related proteins: similarity of adducin with actin-binding proteins. *Protein Seq. Data Anal.*, **1992**, *5*(1), 39-42.

[7] Tripodi, G.; Piscone, A.; Borsani, G.; Tisminetzky, S.; Salardi, S.; Sidoli, A.; James, P.; Pongor, S.; Bianchi, G.; Baralle, F.E. Molecular cloning of an adducin-like protein: evidence of a polymorphism in the normotensive and hypertensive rats of the Milan strain. *Biochem. Biophys. Res. Commun.*, **1991**, *177*(3), 939-947.

[8] Hegyi, H.; Pongor, S. Predicting potential domain homologies from FASTA search results. *Comput. Appl. Biosci.*, **1993**, *9*(3), 371-372.

[9] Murvai, J.; Vlahovicek, K.; Szepesvari, C.; Pongor, S. Prediction of protein functional domains from sequences using artificial neural networks. *Genome Res.*, **2001**, *11*(8), 1410-1417.

[10] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J Mol. Biol.*, **1990**, *215*(3), 403-410.

[11] Pongor, S.; Skerl, V.; Cserzo, M.; Hatsagi, Z.; Simon, G.; Bevilacqua, V. The SBASE domain library: a collection of annotated protein segments. *Protein Eng.*, **1993**, *6*(4), 391-395.

[12] Vlahovicek, K.; Kajan, L.; Murvai, J.; Hegedus, Z.; Pongor, S. The SBASE domain sequence library, release 10: domain architecture prediction. *Nucleic Acids Res.*, **2003**, *31*(1), 403-405.

[13] Vlahovicek, K.; Kajan, L.; Agoston, V.; Pongor, S. The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines. *Nucleic Acids Res.*, **2005**, *33*, D223-D225.

[14] Pongor, S.; Skerl, V.; Cserzo, M.; Hatsagi, Z.; Simon, G.; Bevilacqua, V. The SBASE protein domain library, release 2.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **1993**, *21*(13), 3111-3115.

[15] Pongor, S.; Hatsagi, Z.; Degtyarenko, K.; Fabian, P.; Skerl, V.; Hegyi, H.; Murvai, J.; Bevilacqua, V. The SBASE protein domain library, release 3.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **1994**, *22*(17), 3610-3615.

[16] Murvai, J.; Gabrielian, A.; Fabian, P.; Hatsagi, Z.; Degtyarenko, K.; Hegyi, H.; Pongor, S. The SBASE protein domain library, Release 4.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **1996**, *24*(1), 210-213.

[17] Fabian, P.; Murvai, J.; Hatsagi, Z.; Vlahovicek, K.; Hegyi, H.; Pongor, S. The SBASE protein domain library, release 5.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **1997**, *25*(1), 240-243.

[18] Murvai, J.; Vlahovicek, K.; Barta, E.; Szepesvari, C.; Acatrinei, C.; Pongor, S. The SBASE protein domain library, release 6.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **1999**, *27*(1), 257-259.

[19] Murvai, J.; Vlahovicek, K.; Barta, E.; Cataletto, B.; Pongor, S. The SBASE protein domain library, release 7.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **2000**, *28*(1), 260-262.

[20] Murvai, J.; Vlahovicek, K.; Barta, E.; Pongor, S. The SBASE protein domain library, release 8.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **2001**, *29*(1), 58-60.

[21] Vlahovicek, K.; Murvai, J.; Barta, E.; Pongor, S. The SBASE protein domain library, release 9.0: an online resource for protein domain identification. *Nucleic Acids Res.*, **2002**, *30*(1), 273-275.

[22] Murvai, J.; Vlahovicek, K.; Pongor, S. A simple probabilistic scoring method for protein domain identification. *Bioinformatics*, **2000**, *16*(12), 1155-1156.

[23] Murvai, J.; Vlahovicek, K.; Pongor, S. Towards a memory-based interpretation of proteome data. In: *Supramolecular Structure and Function 7*, Pifat-Mrzljak, G., Ed.; Kluwer Academic Publishers: Dordrecht, **2001**; pp 155-166.

[24] Finn, R.D.; Tate, J.; Mistry, J.; Coggill, P.C.; Sammut, S.J.; Hotz, H.R.; Ceric, G.; Forslund, K.; Eddy, S.R.; Sonnhammer, E.L.; Bateman, A. The Pfam protein families database. *Nucleic Acids Res.*, **2008**, *36*, D281-8.

[25] Letunic, I.; Doerks, T.; Bork, P. SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **2009**, *37*, D229-D232.

[26] Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.C.; Estreicher, A.; Gasteiger, E.; Martin, M.J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **2003**, *31*(1), 365-370.

[27] Stoesser, G.; Baker, W.; van den Broek, A.; Camon, E.; Garcia-Pastor, M.; Kanz, C.; Kulikova, T.; Leinonen, R.; Lin, Q.; Lombard, V.; Lopez, R.; Redaschi, N.; Stoehr, P.; Tuli, M.A.; Tzouvara, K.; Vaughan, R. The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **2002**, *30*(1), 21-26.

[28] Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Wheeler, D.L. GenBank. *Nucleic Acids Res.*, **2005**, *33*, D34-D38.

[29] Jaakkola, T.; Diekhans, M.; Haussler, D. *Using the Fisher kernel method to detect remote protein homologies.* In: Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology, **1999**; pp. 149-58.

[30] Liao, L.; Noble, W.S. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J. Comput. Biol.*, **2003**, *10*(6), 857-868.

[31] Kertesz-Farkas, A.; Dhir, S.; Sonego, P.; Pacurar, M.; Netoteia, S.; Niyween, H.; Kuzniar, A.; Leunissen, J.A.; Kocsor, A.; Pongor, S. A comparison of random and supervised cross-validation strategies and benchmark datasets for protein classification. *J. Biochem. Biophy. Methods*, **2007**, *35*, 1215-1223.

[32] Sonego, P.; Pacurar, M.; Dhir, S.; Kertesz-Farkas, A.; Kocsor, A.; Gaspari, Z.; Leunissen, J.A.; Pongor, S. A Protein Classification Benchmark collection for machine learning. *Nucleic Acids Res.*, **2007**, *35*, D232-D236.

[33] Sonego, P.; Kocsor, A.; Pongor, S. ROC analysis: applications to the classification of biological sequences and 3D structures. *Brief. Bioinform.*, **2008**, *9*(3), 198-209.

[34] Gribskov, M.; Robinson, N.L. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **1996**, *20*(1), 25-33.

[35] Busa-Fekete, R.; Kertesz-Farkas, A.; Kocsor, A.; Pongor, S. Balanced ROC analysis (BAROC) protocol for the evaluation of protein similarities. *J. Biochem. Biophys. Methods*, **2007**, *70*(6), 1210-1214.

[36] Eisenhaber, F. Prediction of Protein Function: Two Basic Concepts and One Practical Recipe. In: *Discovering Biomolecular Mechanisms with Computational Biology*, Eisenhaber, F., Ed. Landes Biosciences Gergetown, **2006**; pp 39-54.

[37] Ooi, H. S.; Kwo, C. Y.; Wildpaner, M.; Sirota, F. L.; Eisenhaber, B.; Maurer-Stroh, S.; Wong, W. C.; Schleiffer, A.; Eisenhaber, F.; Schneider, G. ANNIE: integrated de novo protein sequence annotation. *Nucleic Acids Res.*, **2009**, *37*(*Web Server issue*), W435-40.

[38] Eisenhaber, F. Prediction of Protein Function: Two Basic Concepts and One Practical Recipe. In: *Discovering Biomolecular Mechanisms with Computational Biology*, Springer US: **2006**; pp. 39-54.

[39] Ivanov, D.; Schleiffer, A.; Eisenhaber, F.; Mechtler, K.; Haering, C.H.; Nasmyth, K. Eco1 is a novel acetyltransferase that can acetylate proteins involved in cohesion. *Curr. Biol.*, **2002**, *12*(4), 323-328.

[40] Rea, S.; Eisenhaber, F.; O'Carroll, D.; Strahl, B.D.; Sun, Z.W.; Schmid, M.; Opravil, S.; Mechtler, K.; Ponting, C.P.; Allis, C.D.; Jenuwein, T. Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature*, **2000**, *406*(6796), 593-599.

[41] Schleiffer, A.; Kaitna, S.; Maurer-Stroh, S.; Glotzer, M.; Nasmyth, K.; Eisenhaber, F. Kleisins: a superfamily of bacterial and eukaryotic SMC protein partners. *Mol. Cell*, **2003**, *11*(3), 571-575.

[42] Eisenhaber, B.; Bork, P.; Eisenhaber, F. Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.*, **1999**, *292*(3), 741-758.

[43] Eisenhaber, B.; Schneider, G.; Wildpaner, M.; Eisenhaber, F. A sensitive predictor for potential GPI lipid modification sites in fungal protein sequences and its application to genome-wide studies for *Aspergillus nidulans*, *Candida albicans*, *Neurospora crassa*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. *J. Mol. Biol.*, **2004**, *337*(2), 243-253.

[44] Eisenhaber, B.; Wildpaner, M.; Schultz, C.J.; Borner, G.H.; Dupree, P.; Eisenhaber, F. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. *Plant Physiol.*, **2003**, *133*(4), 1691-701.

[45] Maurer-Stroh, S.; Eisenhaber, B.; Eisenhaber, F. N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J. Mol. Biol.*, **2002**, *317*(4), 541-557.

[46] Maurer-Stroh, S.; Eisenhaber, B.; Eisenhaber, F. N-terminal N-myristoylation of proteins: refinement of the sequence motif and its taxon-specific differences. *J. Mol. Biol.*, **2002**, *317*(4), 523-540.

[47] Maurer-Stroh, S.; Eisenhaber, F. Refinement and prediction of protein prenylation motifs. *Genome Biol.*, **2005**, *6*(6), R55.

[48] Neuberger, G.; Maurer-Stroh, S.; Eisenhaber, B.; Hartig, A.; Eisenhaber, F. Motif refinement of the peroxisomal targeting signal 1 and evaluation of taxon-specific differences. *J. Mol. Biol.*, **2003**, *328*(3), 567-579.

[49] Neuberger, G.; Maurer-Stroh, S.; Eisenhaber, B.; Hartig, A.; Eisenhaber, F. Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*, **2003**, *328*(3), 581-592.

[50] Campagna, D.; Albiero, A.; Bilardi, A.; Caniato, E.; Forcato, C.; Manavski, S.; Vitulo, N.; Valle, G. PASS: a program to align short sequences. *Bioinformatics*, **2009**, *25*(7), 967-968.

[51] Manavski, S.A.; Valle, G. CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. *BMC Bioinformatics*, **2008**, *9*(Suppl 2), S10.

[52] Franklin, D.; Dhir, S.; Pongor, S. Analysis of Kernel Based Protein Classification Strategies Using Pairwise Sequence Alignment Measures. In: *Computational Intelligence Methods for Bioinformatics and Biostatistics*, Springer Berlin / Heidelberg: **2009**; pp. 222-231.

[53] Promponas, V.J.; Enright, A.J.; Tsoka, S.; Kreil, D.P.; Leroy, C.; Hamodrakas, S.; Sander, C.; Ouzounis, C.A. CAST: an iterative algorithm for the complexity analysis of sequence tracts. *Bioinformatics*, **2000**, *16*(10), 915-922.

[54] Dosztanyi, Z.; Csizmok, V.; Tompa, P.; Simon, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **2005**, *347*(4), 827-839.

[55] Brendel, V.; Bucher, P.; Nourbakhsh, I.R.; Blaisdell, B.E.; Karlin, S. Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. USA*, **1992**, *89*(6), 2002-2006.

[56] Wootton, J.C. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **1994**, *18*(3), 269-285.

[57] Cserzo, M.; Eisenhaber, F.; Eisenhaber, B.; Simon, I. TM or not TM: transmembrane protein prediction with low false positive rate using DAS-TMfilter. *Bioinformatics*, **2004**, *20*(1), 136-137.

[58] Tusnady, G.E.; Simon, I. The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **2001**, *17*(9), 849-850.

[59] Kall, L.; Krogh, A.; Sonnhammer, E.L. A combined transmembrane topology and signal peptide prediction method. *J. Mol. Biol.*, **2004**, *338*(5), 1027-1036.

[60] Krogh, A.; Larsson, B.; von Heijne, G.; Sonnhammer, E.L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **2001**, *305*(3), 567-580.

[61] Lupas, A.; Van Dyke, M.; Stock, J. Predicting coiled coils from protein sequences. *Science*, **1991**, *252*(5009), 1162-1164.

[62] Hulo, N.; Bairoch, A.; Bulliard, V.; Cerutti, L.; Cuche, B.A.; de Castro, E.; Lachaize, C.; Langendijk-Genevaux, P.S.; Sigrist, C.J. The 20 years of PROSITE. *Nucleic Acids Res.*, **2008**, *36*, D245-D249.

[63] Eddy, S.R. Profile hidden Markov models. *Bioinformatics*, **1998**, *14*(9), 755-763.

[64] Schaffer, A.A.; Wolf, Y.I.; Ponting, C.P.; Koonin, E.V.; Aravind, L.; Altschul, S.F. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **1999**, *15*(12), 1000-1011.

[65] Marchler-Bauer, A.; Panchenko, A.R.; Shoemaker, B.A.; Thiessen, P.A.; Geer, L.Y.; Bryant, S.H. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **2002**, *30*(1), 281-283.

[66] Ágoston, V.; Kaján, L.; Carugo, O.; Hegedüs, Z.; Vlahovicek, K.; Pongor, S. Concepts of similarity in bioinformatics In: *Essays in Bioinformatics*, Moss, D.S.; Jelaska, S.; Pongor, S.; Eds. IOS Press, Amsterdam: **2005**; pp. 11-31.

[67] Pongor, S. Novel databases for molecular biology. *Nature*, **1988**, *332*(6159), 24.

[68] Pearl, F.; Todd, A.; Sillitoe, I.; Dibley, M.; Redfern, O.; Lewis, T.; Bennett, C.; Marsden, R.; Grant, A.; Lee, D.; Akpor, A.; Maibaum, M.; Harrison, A.; Dallman, T.; Reeves, G.; Diboun, I.; Addou, S.; Lise, S.; Johnston, C.; Sillero, A.; Thornton, J.; Orengo, C. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.*, **2005**, *33*, D247-D251.

[69] Andreeva, A.; Howorth, D.; Chandonia, J.M.; Brenner, S.E.; Hubbard, T.J.; Chothia, C.; Murzin, A.G. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **2008**, *36*, D419-D425.

[70] Tatusov, R.L.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Kiryutin, B.; Koonin, E.V.; Krylov, D.M.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; Rao, B. S.; Smirnov, S.; Sverdlov, A.V.; Vasudevan, S.; Wolf, Y.I.; Yin, J.J.; Natale, D.A. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **2003**, *4*, 41.

Received: October 31, 2009 **Revised: January 03, 2010** **Accepted: July 28, 2010**